The Role of Title, Metadata and Abstract in Identifying Clinically Relevant Journal Articles

Dina Demner-Fushman, MD, PhD, Susan Hauser, PhD, George Thoma, PhD
Lister Hill National Center for Biomedical Communications,
National Library of Medicine, NIH, DHHS, Bethesda, MD

Abstract

Access to current clinical information involves searches of bibliographic databases, such as MEDLINE®, and subsequent evaluation of retrieval results for relevance to a specific clinical situation and quality of the reported research. We establish the amount of information that needs to be provided by an information retrieval system to assist healthcare practitioners in identifying clinically relevant information and evaluating its potential strength of evidence. We find 92% of titles informative enough for a practitioner to correctly classify publications as clinical, but not sufficient for classification of research quality. We suggest automatic organization of retrieval results into strength of evidence categories to supplement title-based judgments and provide quick access to the abstracts of the most promising articles. We find information in the abstracts sufficient to identify articles potentially immediately useful for clinical decision support. These findings are important to the design of information retrieval systems supporting small, low-bandwidth handheld computers.

Introduction

Health care practitioners who would like to combine their personal expertise with the most recent clinical findings obtained in clinical studies need the journal literature, but do not have time to read all new publications. For example, physicians trained in epidemiology would need over 600 hours a month to read every new article published in their field [1]. On the other hand, only a small fraction of articles contains information that has a potential to influence clinical decisions [2]. Not surprisingly, practitioners of evidence-based medicine who initially advocated critical appraisal of the original research found in databases of primary literature now recommend secondary sources that summarize the literature [3]. However these valuable secondary sources cannot predict and review all potential clinical questions. Moreover there is a lag between the publication of the original research and its summarization, and between the updates of the summaries in the secondary databases. These factors suggest there is a need for better access to the relevant primary sources in MEDLINE as the database most frequently searched by clinicians [4]. To that end we are researching ways to improve delivery of clinically pertinent MEDLINE citations at the point of service [5], which in many cases implies using small portable computers [6]. The time and size limitations pose two questions: what is the minimal amount of text that a) provides enough information for a practitioner to identify clinically relevant publications; and b) is sufficient to predict the strength of evidence and the potential immediate clinical validity of the article?

The primary goal of this study is to evaluate how informative is the text available to the user in different modes of interaction with a search engine, which amounts to evaluation of the titles, abstracts, and full text of the article. These three modalities represent the least (title), medium (abstract) and maximal (full-text) time, space and clinician's effort requirements. The two-level evaluation involves: a) determination of the clinical vs. non-clinical orientation of the article, and b) evaluation of the strength of evidence. We find the identification of the strength of evidence information to require efforts beyond browsing the article titles and suggest categorization of the articles into the strength of evidence categories based on MEDLINE indexing as a method to reduce the user's time, space and effort requirements.

Background

Although principles of reading and analyzing medical literature are well researched [7, 8] there is no consensus on what papers are of primary interest to practitioners. Recommendations range from deciding whether a paper merits detailed reading based on the design of the methods section and not on the potential impact of the results [7], to appraising literature in terms of patient oriented outcomes [9]. The latter excludes publications with no potential patient-oriented evidence from further consideration, but then grades the strength of evidence in the selected articles in terms of quality, quantity, and consistency of evidence as reflected in the Strength of Recommendation Taxonomy (SORT) [9]. Our evaluation of the potential immediate clinical validity of the journal article follows the SORT approach, i.e. the text is evaluated for the presence of the patient health outcome, and the grade of the strength of evidence.

Our task is to estimate how much of that information can be identified at each step of a typical interaction with online resources. We are not aware of any work on identification of level of granularity and amount of the text sufficient for understanding of the patient outcome implications; however, studies of the use and adequacy of journal article sections provide insights into contributions of these sections to understanding of scientific publications in general. These studies suggest that in order to determine the focus of a paper and its relevance the researchers primarily use the abstract [10]. The question arises whether MEDLINE abstracts provide information adequate to guide users toward articles containing health outcomes. Research of a more general question, namely "Can the data and other information in the abstract be verified in the body of the article" found that in random samples of 44 articles published in 5 major medical journals 18% to 68% of the abstracts contained data inconsistent with or absent from the article's body [11].

It is not clear, however, if these discrepancies are relevant to abstract-based evaluation of the pertinence of the article. For example, a study that focuses on publications that contain patient-oriented outcomes mentions that in most cases a physician can decide whether an article contains patient-oriented outcomes by scanning the abstract's results [2]. The abstract, closely followed by the results and discussion sections, was found to be the best source of information with respect to gene names, diseases and chemicals, and drugs [12]. The title, abstract and the discussion sections of the paper were also found almost equally useful with regard to gene product information [13].

The classification task in the latter study, to decide whether an article contained material of interest to curators of a genetics database, is very similar to the task faced by clinicians while inspecting retrieval results. Clinician's actions, such as selection of a title to view the abstract, and following the link from the abstract to the full text are in essence classification decisions. This paper describes our work to establish a benchmark for the potential of title and abstract text to guide the clinician to decision support information in the full text.

Methods

We evaluated the amount of information provided in the three typical forms of viewable text using PubMed and search strategies described below to retrieve MEDLINE abstracts and full text of the selected articles.

1. Title evaluation

Abstracts for the title evaluation were retrieved emulating behavior of a user interested in all recent publications on a certain topic. We chose diabetes as the topic of our search, and restricted search results to two weeks time interval ending on the day the search was performed. This search strategy retrieved 632 articles that have been indexed for Medline.

An experienced nurse viewed an onscreen list of the titles without having access to the abstracts and indexing information, and assigned each title to one of the three groups: 1) potentially consistent and good-quality patient-oriented evidence (recommendations strength A [9]); 2) clinically relevant information (recommendations strength B and C[9]), and 3) other. The accuracy of the title-based judgment was evaluated against the gold standard created using the abstracts of the articles that provided the titles.

1.1 Gold Standard

MEDLINE indexing of the 632 articles serves as the gold standard for our evaluation. The gold standard is based on mappings of the types of studies and their ratings used in the strength of evidence taxonomy [9] to the MeSH descriptors and Publication Types assigned by indexers.

The articles were automatically categorized into three groups based on MeSH and Publication Type indexing. The categorization rules are presented in Table 1.

Table 1. ***Publication Type and MeSH-based strength of evidence categories***

1: Clinically Relevant Summary

- Meta-Analysis, Practice Guidelines, Consensus Development Conferences

Clinical Trials

- Controlled Clinical Trials, Randomized Controlled Trials, Multicenter Studies, Double-Blind Method

2: Clinical Evidence

- Studies: Case-Control, Cohort, Cross-Sectional, Cross-Over, Evaluation, Follow-up, Longitudinal, Retrospective, Twin, Validation, Case Reports

Review (not systematic)

3: Other

- Journal Article, Editorial, Interview, Letter, Legal Case
- In Vitro, Animal and Animal Testing Alternatives studies

***2. Metadata evaluation*.**

This experiment explored correlation between the metadata, such as controlled vocabulary terms used in indexing of the article, and potential immediate clinical validity of MEDLINE citations. We studied the following potential validity indicators: publication type of the article; MeSH indexing containing word *outcome;* or presence of the word in the title or abstract. Three annotators evaluated 60 abstracts related to septic shock and labeled each as potentially immediately clinically valid, clinically relevant, or not relevant. Correlation between metadata and the immediate validity of the citation was measured using an asymmetric association statistic, lambda [14], that assesses relative decrease in unpredictability of one variable (e.g., presence of an outcome, or strength of evidence in the citation) when the other variable (e.g. Publication Type, or MeSH descriptor 'Treatment Outcome') is known.

***3. Abstract and full-text evaluation*.**

The goal of the abstract annotation was to evaluate if the amount of information in the abstract is sufficient to a) identify patient outcome information and b) identify the strength of evidence category of the article.

Since articles containing patient oriented outcomes are scarce (2.6% of the publications in 85 medical journals over 6 months according to [2]), we attempted to retrieve only the outcome oriented papers using PubMed and the following search strategy: we selected randomized clinical trials (publication type) with "Treatment Outcome" MeSH descriptor limited to "only items with abstracts", English language, and Humans filters, and providing access to free full text. Outcome statements were then annotated in each of the 137 retrieved abstracts by the first two authors of this paper independently, with good agreement (kappa = 0.75 [14]). After the subsequent reconciliation of differences we excluded two truncated abstracts and classified the remaining 135 abstracts into three groups: without outcomes, with unsupported outcomes, i.e. with no evidence in the abstract that supports the conclusion, and with strongly supported outcomes. We then read 21 articles: all seven that had no outcome statement in the abstract and seven each randomly selected from the other two groups. The randomly selected articles come from the following journals: *Antimicrobial Agents and Chemotherapy*, *Archives of Disease in Childhood, Archives of Physical Medicine and Rehabilitation, BMJ, CMAJ, Chest, Circulation, Health Technology Assessment, JAMA, Journal of the National Cancer Institute,* and *Rheumatology*. We judged the full text of the articles as follows: if the abstract contained clearly stated patient outcome and enough adequate information to tentatively assign it to group A, we looked for discrepancies between the abstract and the other sections of the article that would lead us to reject the evidence in the abstract. For the unsupported statements we attempted to find clearly stated randomization, allocation, and intervention blinding methods, the size of the study, adequacy of the statistical analysis, and the follow-up analysis. For the abstracts without a clear outcome statement we searched for the statement in addition to the above features.

Results

1 Title evaluation

In most cases information content of the titles was sufficient to classify the articles into clinical or not. The annotator achieved 92.2% recall and 93.7% precision. 578 out of the 632 titles were assigned correctly (354 into categories 1 and 2, i.e. clinical, and 224 into the category other). Errors in title evaluation were distributed equally: 24 titles were incorrectly identified as clinical and 30 titles were incorrectly evaluated as not clinical. Title based evaluation of the potential strength of evidence category showed this prediction to be not very reliable. 140 (40%) out of the 354 articles were annotated with the incorrect evidence level.

2. Correlation between metadata and immediate clinical value of the article

Inter-annotator agreement in this experiment was moderate (kappa = 0.55). Metadata is a moderate predictor of the potential clinical value of the citation. Errors in predicting clinical value are reduced by 18 to 32 percent if the publication type is known. Presence of the word *outcome* in MeSH indexing can potentially reduce the prediction error in 23% of the abstracts according to one of the annotators.

3. Correlation between abstracts and full text articles

Our agreement on annotation of the strength of evidence presented in the abstract was very good (kappa = 1.0), and we had no differences in full text evaluation. We found 7 abstracts (5.2%) without outcomes, 37 abstracts (27.4%) with unsupported outcomes, and 91 abstracts (67.4%) with strongly supported outcomes. There was no correlation between the journal and the strength of the outcome support, and the distribution of both structured and un-structured abstracts was similar among the three identified abstract types. There were no discrepancies between the presence and support for the outcome statement in the abstract and in the full text of the article. (See Table 2)

Table 2. ***Abstract outcome statement as a predictor of strength of evidence***

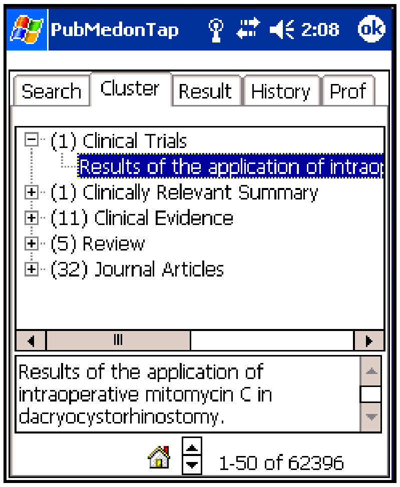
Abstract (7 each)	Full Text Article contains:	
	Patient Outcome	Adequate Supporting Evidence
No Outcome	57%	0%
Unsupported Outcome	100%	0%
Supported Outcome	100%	100%

Discussion

Overall the titles of the articles are highly representative of the clinical orientation of the citation. Considering the recall and precision achieved by the annotator as a measure of the discriminating power of the titles we see that a small percent of the abstracts have a potential to be incorrectly accessed, and slightly more potentially useful articles are missed. Omission errors (false negative assignments) were caused by non-clinically sounding titles, e.g. “The effect of pioglitazone on peroxisome proliferator-activated receptor-gamma target genes related to lipid storage in vivo.” Errors of commission (false positive identification) were caused both by the clinically sounding titles and the annotator’s background knowledge as for example in the following title “Mild nephrogenic diabetes insipidus caused by Foxa1 deficiency” that leads the annotator to assume diabetes occurred in patients, whereas the articles considers Foxa1-deficient mice as a new model of nephrogenic diabetes insipidus.

Identification of the potential evidence level based on the title is problematic. Our proposed solution for this problem is to group the retrieval results into potential strength of evidence categories (see Table 1). For example, Figure 1 demonstrates the presentation of the search results in our PubMed on Tap application for wireless handheld computers [5], clustered by evidence level.

The error analysis of the title evaluation shows potential inaccuracies in categorization based on publication types, for example, Review, can be used as an indicator of the strength of evidence, but is too broad to be used as clinical orientation indicators. Case-Control studies were strongly associated with genetics studies, and it is not clear at present if statements like “the presence of the 192Arg-allele in the PON1 gene is a genetic risk factor for microangiopathy in Type 2 diabetes mellitus” are of interest to primary care providers. Since the evidence strength taxonomy includes the case-control studies we judged the missed case-control studies as false negatives.



PubMedonTap 2:08 ok

Search | Cluster | Result | History | Prof

⊟ (1) Clinical Trials
Results of the application of intraop
⊞ (1) Clinically Relevant Summary
⊞ (11) Clinical Evidence
⊞ (5) Review
⊞ (32) Journal Articles

Results of the application of intraoperative mitomycin C in dacryocystorhinostomy.

1-50 of 62396

Figure 1. ***Evidence strength categorization in PubMed on Tap application***

The goal of our full-text study was to verify the accuracy of the judgments based on the MEDLINE abstracts. Out of the seven abstracts without stated outcomes four had negative results, i.e., full-text showed that the null hypothesis could not be rejected, e.g., in the study of smoking as a risk factor for myocardial infarction smoking history was found to be irrelevant, but this finding was not stated explicitly. Three trials were not concerned with patient outcomes directly: one dealt with ethical issues, another one evaluated feasibility of a potential diagnostic method, and the last one studied professional training of medical personnel. The seven randomly selected abstracts with unsupported outcome statements either did not provide adequate blinding, randomization and follow-up information in the full-text as well, or the results could not be attributed to the intervention under study, as, for example in the study of the effect of a computerized decision support system, where the system was not used. In the remaining seven articles the support for patient outcomes information was reduced to the necessary minimum in the abstracts, and expanded upon in the full-text, e.g., the full text provided exact randomization methodology for the following *Design* section of the abstract: “DESIGN: Randomised, double blind, placebo controlled trial”, but all abstracts accurately reflected the methodology of the study. The patient outcome statements in the abstract and in the full-text were often paraphrases of each other, as for example in the following full-text: “In summary, following three, weekly sclerosant injections to the lumbar spinal ligaments we have been unable to demonstrate improvement in pain, self-reported function, somatization, depression or spinal flexion in patients with undifferentiated chronic back pain” and the abstract “CONCLUSIONS: Three, weekly sclerosant injections alone may not be effective treatment in many patients with undifferentiated chronic back pain.” statements.

The suggestion that the abstracts are sufficiently indicative of the presence of patient outcomes in the full text of the article [2] was confirmed in our study.

Conclusions and Future Work

Our exploratory study demonstrates the majority of the titles in MEDLINE citations to be informative enough for coarse classification into clinical and not-clinical by an experienced practitioner.

In researching the correlation between patient oriented outcome information in the abstract and in the full text of the article we found that outcome statements strongly supported in the abstracts are a good reflection of the outcomes presented in full text. Full texts for abstracts with unsupported or missing outcome statements are harder to read, and do not provide a concise outcome statement. It is unlikely that such full text will be useful to immediately support a clinical decision.

This knowledge is immediately useful as a design consideration for PubMed on Tap, our application to support online MEDLINE searching from a handheld computer and targeting mobile healthcare professionals. Article titles organized into the strength of evidence categories are reasonable portions of the citation to use for first level decisions about finding immediately useful articles while conserving both bandwidth and display space (Figure 1). We have started exploration of the automatic assignment of the strength of evidence levels to the abstracts for which MeSH indexing is not available.

Although second level decisions generally require information from the abstract, strong outcomes statements in the abstract are a good predictor of immediately useful information in the full text. We can assist the busy practitioner further by automatically identifying those outcomes statements in the abstract, a focus of our ongoing research.

Acknowledgements

We thank Thomas Rindflesch for helpful comments on the draft of this paper and Barbara Few for annotation of the titles.

References

[1] Alper BS, Hand JA, Elliott SG, Kinkade S, Hauan MJ, Onion DK, Sklar BM. How much effort is needed to keep up with the literature relevant for primary care? J Med Libr Assoc 2004;92(4) :429-437

[2] Ebell MH, Barry HC, Slawson DC, Shaughnessy AF. Finding POEMs in the medical literature. J Fam Pract 1999;48:350-355.

[3] White B. Making evidence-based medicine doable in everyday practice. Family Practice Management 2004;11(2):51-58

[4] De Groote SL, Dorsch JL. Measuring Use Patterns of Online Journals and Databases. J Med Libr Assoc. 2003; 91(2): 231–241.

[5] Hauser SE, Demner-Fushman,D, Ford G, Thoma, GR. A Testbed System for Mobile Point-of-Care Information Delivery. The 17th IEEE Symposium on Computer-Based Medical Systems (CBMS 2004).

[6] Chen ES, Mendonca EA, McKnight LK, Stetson PD, Lei J, Cimino JJ. PalmCIS: a wireless handheld application for satisfying clinician information needs. J Am Med Inform Assoc. 11 (2004) 19-28.

[7] Greenhalgh, Trisha How to Read a Paper. The Basics of Evidence Based Medicine, 2nd Ed BMJ Books 2000

[8] Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice. Guyatt GH, Rennie D, editors. Chicago: AMA Press, 2002

[9] Ebell MH, Siwek J, Weiss BD, Woolf SH, Susman JL, Ewigman B, Bowman M. Strength of Recommendation Taxonomy (SORT): A Patient-Centered Approach to Grading Evidence in medical literature. J Fam Pract. 2004 53(2):111-20.

[10] Bishop AP. Digital Libraries and Knowledge Disaggregation: The Use of Journal Article Components. ACM DL 1998: 29-39

[11] Pitkin RM, Branagan MA, Burmeister LF. Accuracy of data in abstracts of published research articles. JAMA. 1999 Mar 24-31;281(12):1110-1.

[12] Shah PK, Perez-Iratxeta C, Bork P, Andrade MA. Information extraction from full text scientific articles: where are the keywords? BMC Bioinformatics 2003 May 29;4(1):20

[13] Sinclair G, Webber B. Classification from Full Text: A Comparison of Canonical Sections of Scientific Papers. ISMB 2004

[14] Siegel S, Castellan NJ, Nonparametric Statistics for the Behavioral Sciences, 2nd Ed. 1988 McGraw-Hill